https://doi.org/10.5392/JKCA.2020.20.09.439

공공데이터 개방표준 데이터의 품질평가

Quality Evaluation of the Open Standard Data

김학래
중앙대학교 사회과학대학 문헌정보학과

Haklae Kim(haklaekim@cau.ac.kr)

요약

공공데이터는 공공기관이 전자적으로 생성 또는 취득하여 관리하고 있는 모든 정보와 전자화된 파일이다. 공공데이터는 인공지능, 스마트 시티 등 차세대 신산업을 견인하는 중요한 요소로 인식되고 있다. 한국은 공공데이터 개방과 관련된 국제 평가에서 연속적으로 높은 순위에 위치하고 있다. 그럼에도 불구하고 공공데이터의 활용과 산업적 영향은 미흡하다. 공공데이터의 활용이 미흡한 이유는 다양할 수 있지만, 데이터 품질은 지속적으로 논의되는 주요 이슈이다. 본 논문은 공공데이터 품질 평가를 위한 지표를 검토하고, 개방된 공공데이터를 대상으로 정량적 품질 평가를 수행한다. 특히, 공공데이터 관리지침을 기준으로 구축 및 개방된 개방표준 데이터의 품질을 진단하여 정부의 가이드라인이 적합한지 검토한다. 데이터 품질평가는 개방표준 데이터의 메타데이터와 데이터값을 포함하고, 완전성과 정확성 지표를 기준으로 검토한다. 데이터 분석결과를 바탕으로 품질 개선을 위한 정책적·기술적 방안을 제안한다.

■ 중심어 : | 공공데이터 | 개방표준 | 데이터품질 | 데이터포털 |

Abstract

Public data refers to all data or information created by public institutions, and public information that leads to communication and cooperation among all people. Public data is an important method to lead the next generation of new industries such as artificial intelligence and smart cities, Korea is continuously ranked high in the international evaluation related to public data. However, despite the continuous efforts, the use of public data or industrial influence is insufficient. Quality issues are continuously discussed in the use of public data, but the criteria for quantitatively evaluating data are insufficient. This paper reviews indicators for public data quality evaluation and performs quantitative evaluation on selected public data. In particular, the quality of open standard data constructed and opened based on public data management guidelines is examined to determine whether government guidelines are appropriate. The data quality assessment includes the metadata and data values of open standard data, and is reviewed based on completeness and accuracy indicators. Based on the data analysis results, this paper proposes policy and technical measures for quality improvement.

■ keyword : | Public Data | Open Standard | Data Quality | Data Portal |

접수일자 : 2020년 06월 26일
수정일자 : 2020년 08월 04일
심사완료일 : 2020년 08월 04일
교신저자 : 김학래, e-mail : haklaekim@cau.ac.kr

I. 서 론

대한민국 정부는 공공데이터의 제공 및 이용활성화에 관한 법률(약칭 : 공공데이터법, 2013년 7월)을 제정하고, 이 법률을 근거로 공공데이터를 개방하고 있다[1]. 공공데이터는 공공기관이 업무 수행의 결과로 생성 또는 취득한 모든 자료를 말하며, 텍스트, 수치, 오디오, 이미지, 동영상 등 다양한 형식의 전자화된 파일을 포함될 수 있다. 공공데이터 개방은 공공기관이 보유·관리하는 공공데이터를 이용자가 자유롭게 활용할 수 있도록 다양한 형태로 제공하는 것이다[2]. 이때, "제공"이란 사용자가 기계 판독이 가능한 형태로 공공데이터에 접근할 수 있거나 다양한 방식으로 전달하는 것을 의미하며, "기계 판독"은 소프트웨어로 데이터의 내용 또는 구조를 확인하고 수정·변환·추출 등 가공할 수 있는 상태를 말한다[2]. 개념적으로 보면, 공공데이터는 오픈 데이터와 유사한 의미를 갖고 있으며, 공공영역 정보 (public sector information) 또는 개방형 정부 데이터 (open government data)와 맥락적으로 일치한다[3].

2020년 6월 현재, 공공데이터포털은 878개 공공기관으로부터 31,814건의 파일 데이터, 5,586 건의 오픈 API (Application Programming Language), 120건의 개방표준 데이터를 개방하고 있다[1]. 한국은 공공데이터와 관련된 국제 평가도구인 OECD (Organisation for Economic Co-operation and Development)의 OUR Data Index[4], Web Foundation의 Open Data Barometer에서 연속적으로 높은 순위로 평가받으며 리더 국가로 자리매김하고 있다. 특히, Open Data Barometer (2018)에서 한국은 프랑스와 함께 공동 4위로 평가되었으며, 지속적인 공공데이터 개방이 사회 혁신을 이끌어가는 정부의 모습이 강조되고 있다[5].

그러나, 지속적인 데이터 개방에도 불구하고 공공데이터 활용은 여전히 미흡하고, 데이터 사용자로부터 비판적인 의견이 존재하는 것이 현실이다[6][7]. 공공데이터는 공공기관이 보유하고 있는 원천시스템에서 일부 데이터가 개방되며, 이 과정에서 데이터가 갖고 있는 의미를 잃을 수 있다. 뿐만 아니라, 개방된 공공데이터의 표현 형식, 데이터 내용에 대한 일관성이 부족하기 때문에 데이터 품질에 대한 이슈가 지속적으로 발생하고 있다[8]. 정부는 공공데이터 품질관리 수준평가를 위해 평가 모델과 지표를 구체화하고[9][10], 공공기관을 대상으로 품질 평가를 위한 실태조사를 진행하고 있다[11]. 그러나, 기관이 아닌 데이터셋 중심의 데이터 품질을 진단하는데 여전히 한계가 있다.

본 논문은 개방표준 데이터의 품질 수준을 진단하고, 이를 바탕으로 공공데이터 품질 개선 방안을 제안한다. 논문의 구성은 다음과 같다. 2장은 데이터 품질평가에 대한 관련 연구를 검토한다. 3장은 개방표준 데이터의 품질 평가 방법과 결과를 요약한다. 마지막으로, 4장에서 논문의 공헌점과 향후 연구에 대해 기술한다.

II. 관련 연구

데이터 품질은 학계와 산업계에서 광범위하게 연구되고 있다[12-14]. 데이터 품질의 측정 방법은 다양한 해석과 방법이 존재한다. 일반적으로, 데이터 품질은 정확성 (accuracy), 완전성 (completeness), 일관성 (consistency)과 같은 다양한 차원으로 평가된다[12].

오픈 데이터 관점에서 데이터 개방 규모만큼 양질의 데이터를 제공하는 것이 중요하다. 최근 여러 연구에서 오픈 데이터의 품질을 분석했으며[15], 데이터 품질의 측정 기준과 항목과 관련한 한계를 지적하고 있다[12]. 열린지식재단은 오픈 데이터의 품질이 일회성 측정이 아닌 프로세스 차원에서 지속적으로 관리되는 것이 중요하다고 지적한다[16].

국가별로 운영되는 포털의 데이터 품질을 분석한 연구도 다양하게 진행되고 있다[17]. Global Open Data Index[2]와 Open Data Barometer[3]는 세계 각국 정부가 개방하는 데이터를 주제별로 구분해 데이터 품질을 평가하는 도구로 활용되고 있고, 분석 결과는 매해 데이터 지표로 변환해서 공유하고 있다. Viscusi et.al. [18]은 이탈리아의 공공데이터를 데이터 제공주체 (지역, 지방자치단체)별로 구분하여 완전성, 정확성

1 http://data.go.kr

2 https://index.okfn.org/

3 https://opendatabarometer.org/

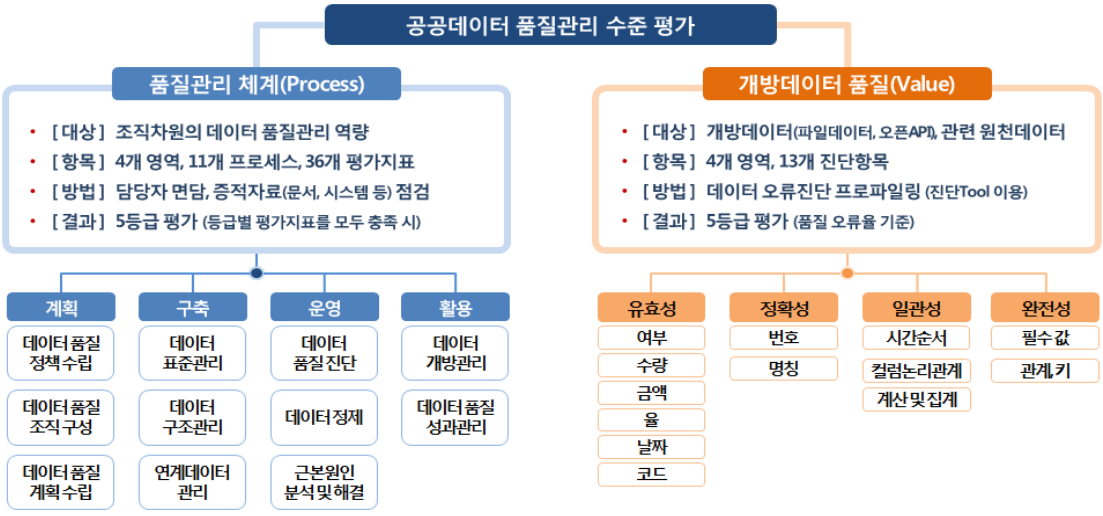

그림 1. 공공데이터 품질관리 수준평가 (공공데이터 관리지침, 2017)

및 적시성 측면에서 품질을 평가하고 있다. 개별 데이터집합에서 계산된 완전성, 정확성은 데이터 포털에서 집계하고 데이터 품질 수준을 측정한다. 분석 결과에 의하면, 지방자치단체가 운영하는 포털의 40%가 완전하지 않은 데이터를 제공하고, 기계가 읽을 수 없는 형식의 데이터셋이 55% 이상으로 보고하고 있다. Tauberer[19]는 미국 의회 구성원에 대한 데이터가 완전히 다른 스키마와 ID로 표현되어 있어, 데이터 통합이 어렵다는 결과를 보고하고, 데이터 품질 개선이 오픈데이터 활용에 필수적이라고 주장한다. 김학래 [8]는 한국의 공공데이터포털에 개방된 데이터에서 메타데이터를 추출하고, 필드명 수준의 품질을 평가하고 있다. 추출된 필드명은 한글뿐만 아니라 다국어가 포함되어 있고, 일관되지 않은 필드명이 많기 때문에 데이터의 연계에 있어 한계가 존재한다고 지적한다.

그러나 공공데이터 품질 평가는 메타데이터 중심으로 평가되고 있기 때문에 데이터셋이 갖고 있는 실제 데이터값의 품질도 평가가 필요하다. 본 논문은 개방표준 데이터에 포함된 데이터셋에서 데이터값을 추출하고, 개별 데이터의 값을 진단해 개별 데이터셋의 종합적인 품질평가를 수행한다.

III. 공공데이터 품질관리

1. 개요

공공데이터법 제22조(공공데이터의 품질관리)는 공공데이터의 안정적 품질관리 및 적정한 품질수준의 확보를 위하여 품질 진단·평가, 개선 지원 등 필요한 시책을 수립하고, 사회적·경제적 파급효과가 큰 공공데이터에 대한 품질 평가를 실시하고 결과를 공표할 수 있다고 규정하고 있다[2]. 이에 따라 정부는 공공데이터 품질관리 수준평가 모델을 개발하고, 2016년부터 공공데이터 품질관리 수준평가를 시행하고 있다. 2016년과 2017년에 실시한 수준평가는 각각 21개, 42개 데이터베이스를 선정해 시행되었고, 2018년부터 중앙행정기관, 지방자치단체, 공공기관 등 600여개 기관을 대상으로 연차적으로 확대할 계획이다[11].

공공데이터 품질관리 수준평가는 품질관리 체계와 개방데이터 품질로 구분된다[10]. 전자는 조직 차원에서 데이터 품질을 관리할 수 있는 역량을 측정하며 계획·구축·운영·활용 등 4개 영역의 11개 프로세스를 포함한다. 계획과 구축 영역은 각각 품질 관리를 위한 계획과 정책 수립, 데이터 표준과 구조 관리를 평가지표로 설정하고 있고, 운영과 활용 영역은 데이터 품질 진단과 개방 및 성과 관리 요소를 포함하고 있다. 개방데이터 품질은 파일데이터, 오픈 API 등 공공데이터로 개방된 데이터의 유효성, 정확성, 일관성, 완전성을 평가한다. 13개 항목으로 구분된 진단항목은 데이터값의 구

표 1. 품질관리를 위한 평가등급 정의 (행정안전부, 2018)

구분	설 명
1등급 (최적화)	조직 전체의 데이터 품질관리 활동의 선순환체계가 확립되고, 이를 통해 공공데이터의 안정적 품질향상 및 유지가 보장되는 수준
2등급 (체계화)	조직 차원의 데이터 품질관리 프로세스가 이행되고, 데이터 품질관리 활동이 체계적 수행이 가능한 수준
3등급 (관리화)	데이터 품질관리를 위한 필수적인 활동들이 관리 및 통제되어, 이를 통해 데이터 품질 향상이 가능한 수준
4등급 (도입)	데이터 품질관리가 인식되고, 품질진단 등 기초적인 품질관리 활동들이 도입시작하는 수준
5등급 (도입전)	데이터 품질관리가 인식이 미흡하여 기본적인 품질관리 활동의 수행이 불가능하거나 부분적인 품질관리 활동만 수행되는 수준

조적 형식을 진단하는데 초점이 있다.

지표별 점수는 평정 척도 (rating scale)의 한 종류인 리커트 척도(Likert scale)를 이용한다. 이 척도는 하나의 주제를 문구 또는 문장으로 제시하고 응답자가 응답한 전반적인 경향을 측정치로 합산하여 결과 점수를 도출하는 방법이다. 계획영역은 기관 차원의 평가지표로 총합이 20점이며, 지표의 충족 여부에 따라 5점 척도의 만족도를 측정하고 그 합을 영역점수로 할당한다. 반면, 구축·활용 영역은 기관이 보유하고 있는 데이터베이스를 진단하기 위한 지표로 구성되고, 평가점수의 합산 평균을 지표점수로 산정한다. 마지막으로 종합점수는 각 영역별 점수를 합산하여 산출하고 등급기준을 적용해 기관 최종등급을 부여한다. 품질관리 수준평가의 결과는 도입전 → 도입 → 관리화 → 체계화 → 최적화의 5등급 체계로 구분하며, 등급별 점수분포는 1등급(100~81점), 2등급(80~61점), 3등급(60~41점), 4등급(40~21점), 5등급(20점 이하)이다.

2. 공공데이터 품질관리 수준 평가의 한계

공공데이터 품질관리 수준평가는 범정부 차원에서 시행되고 있지만, 공공데이터 관점으로 보면 평가 대상과 목표가 명확하지 않다. 첫째, 수준 평가는 기관 고유의 행정업무 수행을 위하여 생성, 취득하여 운영하는 모든 데이터베이스를 대상으로 하며, 영향도, 연계규모, 이용자 활용도 관점에서 점수가 높은 상위 25%를 대상으로 선정한다. 그러나 평가대상 데이터베이스가 공공데이터로 개방되었는지 여부와 관계없이 선정되기 때문에, 평가결과를 공공데이터의 품질평가로 해석하는데 한계가 있다. 둘째, 공공데이터 품질관리 평가체계의 지표별 점수는 평정 척도(Rating scale)의 한 종류인 리커트 척도(Likert scale)를 이용한다. 통계적 관점으로 보면, 리커트 척도를 활용한 수준 평가는 각 문항에 대한 주관적 판단의 위험이 있고, 총점을 계산하는 과정에서 각 항목에 대한 응답점수의 편차가 사라지기 때문에 총점의 개념적 의미를 명확히 정의하기 어렵다. 실제, 개별 평가지표는 다양한 세부지표를 포함하고 있기 때문에 척도 기반의 점수를 부여하는 것이 효과적이지 않다. 마지막으로, 공공데이터 품질 평가의 주요 대상이 기관과 기관이 보유하고 있는 데이터베이스이기 때문에 개별 데이터 세트에 포함된 메타데이터와 데이터 값을 진단하는데 한계가 있다. [그림 1]의 개방데이터의 품질에 정의된 4개 영역을 보면, 일관성의 컬럼논리관계, 완전성의 키, 관계는 공공데이터의 개방 과정에서 손실되거나, 개방 내용에 포함되어 있지 않다. 더불어, 유효성의 평가항목인 수량, 금액, 코드는 공공데이터에서 참조할 수 있는 대상이 거의 없기 때문에, 실효적인 평가가 어렵다. 이런 이유로 공공데이터의 품질은 실제 데이터에 적용할 수 있는 평가 항목으로 재구성하는 것이 필요하다.

IV. 개방표준 공공데이터의 품질평가

본 연구는 공공데이터 품질평가를 위해 개방 표준데이터를 수집하고, 완전성과 정확성에 대한 평가를 수행한다. 먼저 개방표준 데이터의 현황과 수집 방법을 소

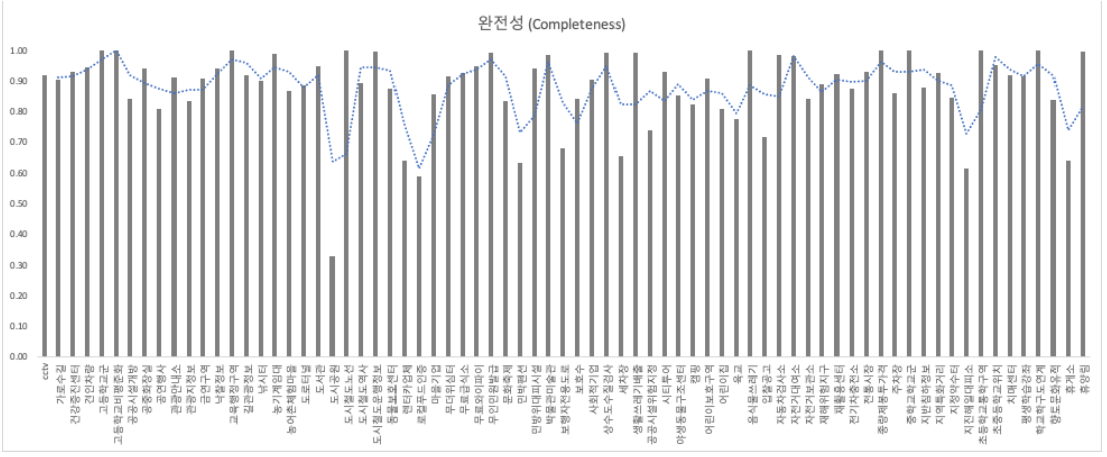

그림 2. 완전성 평가 결과

개하고, 데이터 품질 평가에 대해 요약한다.

1. 데이터 수집

개방 표준 데이터는 행정자치부에서 고시한 공공데이터 개방 표준(제2016-42호)을 근거로 제정한 데이터 세트이며, 데이터 파일형식, 명명규칙, 분야별 제공항목 및 속성정보(표현형식/단위 등)를 표준화하고 있다[20]. 더불어, 개방 표준으로 제정된 분야는 메타데이터와 실제 데이터 값이 포함된 데이터 세트를 제공하고 있다. 이런 특징은 데이터 품질 평가에 있어 매우 중요하다. 즉, 데이터에 대한 품질 측정을 위한 표준 지침이 명시화되어 있고, 동시에 지침을 반영해 구축한 데이터가 존재하기 때문에 데이터 세트의 품질 평가 기준이 명확해진다. 법률에 명시되지 않았지만, 표준 데이터는 가장 바람직하고 정확한 데이터라는 상징적 의미를 갖고 있다.

개방 표준 데이터는 2014년에 주차장, 도시공원으로 시작해 4차 개정을 통해 2018년 10월 현재 총 91건이 정의되어 있다. 전체 데이터셋에서 79건은 csv, json, xml 등의 개방형 파일형식으로, 12건은 오픈 API 형식으로 제공한다. 개별 데이터 세트의 속성 이름 및 데이터 행의 개수는 다양하다. 예컨대, 개방 표준 데이터는 평균적으로 20개의 열과 10,361개의 행을 갖고 있다. 이 중에서 행의 개수는 최소 21개에서 최대 189,962개로 큰 차이가 있었고, 열의 개수도 최소 7개에서 최대 67개로 분포가 다양하다.

2. 데이터 품질 평가

오픈데이터 평가지표는 연구자에 따라 차이가 있지만, 일반적으로 완전성, 정확성, 일관성, 적시성, 보안성을 범용적으로 활용한다[15]. 그러나 개방표준 데이터에 모든 평가 지표를 적용하는데 현실적 제약이 있다. 예컨대, 데이터의 중복 및 정합성을 판단하는 지표인 일관성은 개방 표준 데이터에 포함된 정확한 값을 특정하지 못하기 때문에 측정이 어렵다. 반면, 데이터베이스 성능, 수요 중심의 데이터를 판단하는 적시성과 데이터 접근에 대한 지표인 보안성은 이미 개방되어 있는 데이터에 적용하는데 적합하지 않다. 완전성의 평가항목인 키, 관계는 대부분의 공공데이터에 포함되지 않은 정보이다. 이런 이유로 본 연구는 완전성(Completeness)과 정확성(Accuracy)을 평가 기준으로 정의한다. 다만, 유효성의 품질평가 항목은 정확성과 결합하여 측정한다. 품질평가는 오픈리파인[4]을 이용하고, 진단을 위한 스크립트를 정의해 자동으로 검출한다.

먼저, 완전성은 데이터 세트의 특정 셀이 공백인 모든 셀의 비율을 측정한다. 완전성 지수 (pcc: Percentage of Complete Cells)는 개별 데이터의 완전성 비율로 수식 1과 같이 정의한다. ic(number incomplete cells)는 공백으로 표현된 불완전한 셀의

4 http://openrefine.org

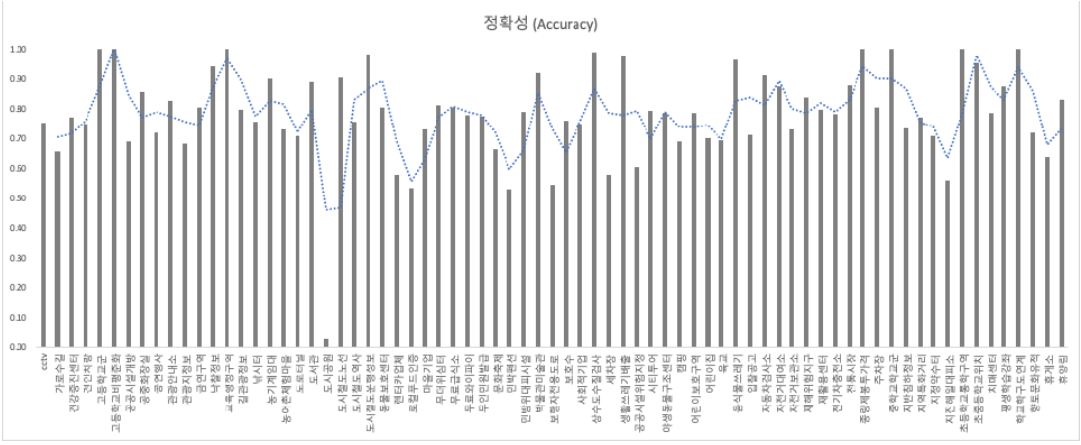

그림 3. 정확성 평가 결과

개수이고, ncl(number of cells)은 전체 셀의 개수를 의미한다.

$$pcc = (1-ic/ncl) \times 100 \text{ ----------(1)}$$

정확성은 유효성의 특징을 포함한 데이터 값의 구문적 오류를 측정한다. 수식 2는 정확성 지수 (pae: Percentage of Accurate Cells)는 나타내며 nce (number of cells with errors)는 오류를 포함한 셀의 개수를 의미한다.

$$pae = (1-nce/ncl) \times 100 \text{ ----------(2)}$$

마지막으로, 품질상관계수 (Quality correlation coefficient)는 완전성과 정확성 지표의 차이로 계산한다. 특정 데이터의 품질상관계수가 크면, 데이터 품질이 낮아진다.

$$qcc = (pcc - pae) \text{ ----------(3)}$$

3. 결과

완전성과 정확성 측면에서 개방표준 데이터에 대한 품질 평가는 비교적 높은 수준이다. 완전성 지표의 평균은 0.88이다. 개방 표준 데이터는 표준화 대상으로 선정되면, 데이터 제공 기관 또는 지방자치단체에서 개방한 데이터를 통합하는 방식으로 데이터를 구축된다. 이런 과정에서 데이터 세트의 개별 필드를 구성하는 값이 누락될 수 있다. 실제 개방표준 데이터에 존재하는 오류 유형 중 공백 오류가 차지하는 비중은 66.6%로 가장 높다 (특수문자 오류 5%, 구문 오류 19.4%, 값 오류 9.1%). 특히, '도시공원' 개방 표준 데이터의 완전성 지수는 0.33으로 가장 낮았다. 개방표준 데이터는 공공데이터의 모범사례로 구축되었지만, 품질 관리 측면에서 완전성 항목은 보완이 필요하다. 이런 결과는 정확성 지수에도 영향을 미친다. 완전성 지수가 낮은 상위 10개 데이터는 '도시공원', '로컬 푸드 인증', '지진해일 대피소', '민박펜션', '렌터카업체', '휴게소', '세차장', '보행자전용도로', '입찰공고', '공공시설위험지정'이 포함된다[그림 2].

정확성의 평균은 0.78로 완전성에 비해 낮게 평가되었다. 정확성은 필드에 포함된 값이 측정 대상이고, 정확한 값으로 표현되었는지 여부를 판단한다. 다만, 개별 데이터셋에 있는 데이터의 정답, 즉 참조값을 특정하기 어렵기 때문에 의미적 정확성을 판단하는 것은 어렵다. 따라서, 정확성은 값이 가지는 의미(semantics)보다는 구문(syntax)적 특성을 측정하였다. '고등학교군', '고등학교비평준화', '교육행정구역'은 정확성 지수가 높게 나타난다. 해당 데이터 세트는 공공데이터로의 개방 여부와 관계없이 해당 도메인에서 표준으로 사용되고 있거나, 데이터의 정제를 통해 오류가 매우 적다는 특징이 있다. 반면, '도시공원', '로컬 푸드 인증', '민박펜

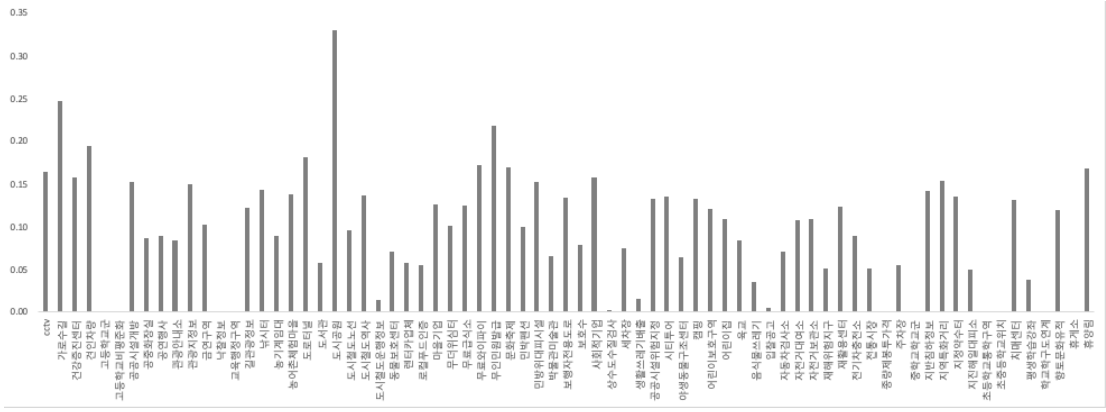

그림 4. 품질상관계수 평가 결과

션', '보행자전용도로', '지진해일대피소', '렌터카업체', '세차장', '공공시설위험지정', '휴게소', '가로수길' 등 완전성 지수가 낮은 개방 표준은 정확성 지수도 낮다[그림 3]. 정확성에 영향을 미치는 오류는 특수문자 오류, 구문 오류, 값 오류가 있다. 특수문자는 데이터를 구축 또는 변환하는 단계에서 다양한 특수문자가 포함되는 것으로, 기존의 파일 형태의 문서를 데이터로 변환하는 과정에서 발생할 수 있다. 구문 오류는 날짜, 주소, 수량 등 일관성 있는 표기법을 지키지 않을 때 발생하는 오류이며, 개방 표준 데이터에서 광범위하게 나타나고 있다. 특히, 위경도 좌표를 포함하는 위치정보의 경우 정확성 지수가 낮은 경우가 많은데, 위경도의 표현 방법이 일관적이지 않아 발생된다. 마지막으로 값 오류는 잘못된 데이터 값이 포함된 경우이다. 예를 들어, 관리기관명에 대한 값은 '서울특별시'가 아닌 '서울특별시청'이 올바르지만, 이처럼 표준 명칭을 따르지 않는 오류가 높게 발생되고 있다. 마지막으로, 그림 4에서 보듯이, 완전성과 정확성의 차이인 품질상관계수는 '고등학교군', '낙찰정보', '교육행정교육', '휴게소' 등 10개 데이터의 지표값이 0으로 나타난다. 반면, '가로수길', '도시공원', '무인민원발급'은 각각 0.25, 0.33, 0.22로 품질상관계수의 값이 다른 데이터와 비교해 상대적으로 높은 것을 볼 수 있다. 즉, 해당 데이터는 품질 개선이 필요하다고 해석할 수 있다.

요약하면, 개방 표준 데이터에 대한 완전성과 정확성 지수는 각각 0.88, 0.78로 비교적 높은 수준이다. 그러나, 국가에서 인증한 표준 데이터는 더욱 높은 수준의 품질로 제공하는 것이 필요하다. 완전성이 낮은 데이터 셋의품질을 우선 관리하여 개방 표준 데이터의 전반적인 품질을 향상시킬 필요가 있다. 더불어 공공데이터에서 보편적으로 활용되는 '위치', '관리기관', '주소' 정보는 정확성을 개선하기 위해 우선적으로 고려해야 한다.

V. 결론

본 논문은 공공데이터 품질관리에 대한 선행 연구를 고찰하고, 개방표준 데이터를 대상으로 품질 평가를 수행하였다. 데이터 품질은 데이터 관리를 위해 중요한 요소이며, 일반적으로 데이터베이스의 구축과 관리에서 명시적인 지침을 마련하고 있다. 그러나 공공데이터의 품질은 기존의 이론적 틀로 해석하는데 한계가 있기 때문에, 이에 대한 품질 평가는 체계적으로 진행되지 못하고 있는 것이 현실이다. 본 논문은 공공데이터 관리 지침의 기준에 의해 구축된 개방표준 데이터를 대상으로 품질 평가를 수행했다. 분석 결과에 따르면, 개방표준 데이터는 완결성 측면에서 88% 이상이었지만, 정확성은 상대적으로 낮은 것으로 나타났다. 즉, 개방표준 데이터는 국가의 표준에 의해 구축되었으나, 제공 데이터의 품질개선이 필요하다.

공공데이터 품질 개선은 제도적·기술적 관점으로 구분해 검토해야 한다. 첫째, 공공데이터 품질관리에 대한

개념화가 필요하다. 전통적인 데이터 품질관리는 데이터 계획과 구축 및 관리 영역을 체계화하고 있지만, 데이터의 개방과 활용에 대한 이론적 틀은 미흡하다. 둘째, 공공데이터 관련 법제도와 관리지침이 공공데이터의 특성을 반영한 품질 요소를 반영하는 것이 중요하다. 공공데이터 관리지침, 개방표준 데이터 및 공공데이터 품질평가는 정책적으로 필요하지만, 외부에 개방된 공공데이터의 특성을 반영해야 한다. 실제, 개방데이터 품질의 4개 영역은 일반적인 데이터 품질 평가 항목으로 적합할 수 있으나, 공공데이터의 특성을 반영해 세부 평가 지표를 정의해야 한다. 셋째, 정부의 원천시스템에서 공공데이터로 개방되는 단계에 적용할 수 있는 프로세스를 체계화하고, 품질 평가와 개선을 지원할 수 있는 기술적 지원(예: 소프트웨어)이 필요하다.

본 연구는 한정된 데이터 세트를 대상으로 완전성과 정확성 지표로 데이터 품질을 평가했기 때문에 공공데이터 전반으로 해석하는데 한계가 있다. 공공데이터는 기존의 데이터베이스와 다른 특성이 있어 데이터 품질 평가를 위해 새로운 지침이 필요하다. 향후 연구는 공공데이터포털에 개방된 전체 데이터 세트를 평가할 수 있는 프레임워크를 개발하고, 메타데이터와 더불어 데이터 값에 대한 품질을 평가하기 위한 방법론을 포함한다.

참 고 문 헌

[1] 행정안전부, "공공데이터의 제공 및 이용 활성화에 관한 법률 (약칭:공공데이터법) 법률 제11956호", 2013.

[2] 행정안전부, "공공데이터 관리지침", 행정자치부/공공데이터활용지원센터, 2017.

[3] 김학래, "공공데이터의 의미적 연계를 위한 행정구역 지식 그래프 구축," 한국콘텐츠학회논문지, Vol.17, No.12, pp.1-10, 2017.

[4] OECD, Government at a Glance 2017, OECD Publishing, Paris, 2017,

[5] Open Data Barometer, https://opendatabarometer.org/, 2018.10.

[6] 강희종, "공공데이터의 효율적 활용을 위한 정책 과제," STEPI Insight, Vol.156, pp.1-30, 2014.

[7] 서형준, "공공데이터 개방에 관한 실증연구: ODB와 OUR Index를 중심으로," NIA 정보화정책, Vol.24, No.1, pp.48-78, 2017.

[8] H. Kim, "Analysis of standard vocabulary use of the open government data: the case of the public data portal of Korea," Quality and Quantity, Vol.53, pp.1611-1622, 2019.

[9] 한국정보화진흥원, "공공데이터 품질관리 매뉴얼 v.2.0," 2018.

[10] 행정안전부, "공공데이터 품질관리 수준평가 가이드," 2017.

[11] 공공데이터전략위원회, "2017년도 품질관리 수준평가 결과," 2019.

[12] S. W. Sadiq and M. Indulska, "Open data: Quality over quantity," Int J. Information Management, Vol.37, No.3, pp.150-154, 2017.

[13] D. Corsar and P. Edwards, "Challenges of Open Data Quality: More Than Just License, Format, and Customer Support," Journal of Data and Information Quality, Vol.9, No.1, pp.1-3, 2017.

[14] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," ACM computing surveys (CSUR), Vol.41, No.3, p.16, 2009.

[15] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: Definition and application to Open Government Data," Government Information Quarterly, Vol.33, No.2, pp.325-337, 2016.

[16] Open Knowledge Foundation, "The Open Dat a Handbook," https://opendatahandbook.org/

[17] Máchová, Renáta and Lněnička, Martin, "Evaluating the Quality of Open Data Portals on the National Level," Journal of theoretical and applied electronic commerce research, Vol.12, pp.21-41, 2017.

[18] Viscusi Gianluigi, Spahiu Blerina, Maurino Andrea, and Batini Carlo, "Compliance with open government data policies: An empirical assessment of Italian local public administrations," Information Polity, Vol.19,

2014.
[19] Joshua Tauberer, “Open Government Data,” 2012. https://opengovdata.io
[20] 행정안전부, “공공데이터 개방 표준: 행정안전부고시 제 2019-1호,” 2018.

저 자 소 개

김 학 래(Haklae Kim) 정회원

- 2010년 6월 : 아일랜드 국립대학교 (공학박사)
- 2019년 3월 ~ 현재 : 중앙대학교 문헌정보학과 교수

〈관심분야〉 : 지식공학, 인공지능, 데이터 사이언스